Derivation of Two Critical Appraisal Scores for Trainees to Evaluate Online Educational Resources: A METRIQ Study

Teresa M. Chan, MD, MHPE*
Brent Thoma, MD, MA†
Keeth Krishnan, BHSc‡
Michelle Lin, MD§
Christopher R. Carpenter, MD, MSc¶
Matt Astin, MD, MPH‖
Kulamakan Kulasegaram, PhD#**

*McMaster University, Department of Medicine, Division of Emergency Medicine, Hamilton, Ontario, Canada
†University of Saskatchewan, Department of Emergency Medicine, Saskatoon, Saskatchewan, Canada
‡University of Toronto, Faculty of Medicine, Undergraduate Medical Education, Toronto, Ontario, Canada
§University of California, San Francisco, School of Medicine, Department of Emergency Medicine, San Francisco, California
¶Washington University School of Medicine, St. Louis, Missouri
‖Mercer University School of Medicine, Department of Emergency Medicine, Department of Internal Medicine, Macon, Georgia
#University of Toronto Faculty of Medicine, Department of Family and Community Medicine, Toronto, Ontario, Canada
**University Health Network, Department of Emergency Medicine, Wilson Centre for Health Professions Education, Toronto, Ontario, Canada

Section Editor: Mark I. Langdorf, MD, MHPE
Submission history: Submitted May 5, 2016; Revision received June 22, 2016; Accepted June 30, 2016
Electronically published July 26, 2016
Full text available through open access at http://escholarship.org/uc/uciem_westjem
DOI: 10.5811/westjem.2016.6.30825

Introduction: Online education resources (OERs), like blogs and podcasts, increasingly augment or replace traditional medical education resources such as textbooks and lectures. Trainees' ability to evaluate these resources is poor, and few quality assessment aids have been developed to assist them. This study aimed to derive a quality evaluation instrument for this purpose.

Methods: We used a three-phase methodology. In Phase 1, a previously derived list of 151 OER quality indicators was reduced to 13 items using data from published consensus-building studies (of medical educators, expert podcasters, and expert bloggers) and subsequent evaluation by our team. In Phase 2, these 13 items were converted to seven-point Likert scales used by trainee raters (n=40) to evaluate 39 OERs. The reliability and usability of these 13 rating items was determined using responses from trainee raters, and top items were used to create two OER quality evaluation instruments. In Phase 3, these instruments were compared to an external certification process (the ALiEM AIR certification) and the gestalt evaluation of the same 39 blog posts by 20 faculty educators.

Results: Two quality-evaluation instruments were derived with fair inter-rater reliability: the METRIQ-8 Score (Inter class correlation coefficient [ICC]=0.30, p<0.001) and the METRIQ-5 Score (ICC=0.22, p<0.001). Both scores, when calculated using the derivation data, correlated with educator gestalt (Pearson's r=0.35, p=0.03 and r=0.41, p<0.01, respectively) and were related to increased odds of receiving an ALiEM AIR certification (odds ratio=1.28, p=0.03; OR=1.5, p=0.004, respectively).

Conclusion: Two novel scoring instruments with adequate psychometric properties were derived to assist trainees in evaluating OER quality and correlated favourably with gestalt ratings of online educational resources by faculty educators. Further testing is needed to ensure these instruments are accurate when applied by trainees. [West J Emerg Med. 2016;17(5)574-584.]

INTRODUCTION

With widespread access to and use of the Internet, there have increasingly been calls by the academic community for scientists to share their knowledge with the public and data with fellow researchers.[1-2] Consistent with this open access movement, there has been a push to expand the repository of online educational resources (OERs). In medical education, this movement has been dubbed Free Open Access Medical education (FOAM). Social media platforms, such as blogs and podcasts, have catalyzed the proliferation of OERs partly because of their ease of publishing.[3-4] Because these resources are readily accessible and literally at the fingertips of most clinicians and trainees, they are increasingly supplanting both medical journals and textbooks as a leading source of individualized, asynchronous learning.[5-7] Furthermore, healthcare professionals are forming virtual communities of practice to share knowledge and network with their peers and trainees, revolving around these social media platforms.

With these new resources comes the burden of teaching learners and educators how to critically appraise them. Just as critical appraisal of primary literature is a key component of a robust medical education, so too is the ability to critically read secondary reference materials such as review papers and textbooks. However, whereas most medical school and residency curricula are required to incorporate the critical appraisal of the medical literature,[8-9] little attention is given to appraising secondary resources such as textbooks, lectures, and OERs. This is concerning because inter-rater reliability of gestalt ratings of these products by trainees is quite poor.[10] Whereas multiple critical appraisal instruments have been published to assist clinicians in the evaluation of the literature (e.g. the *Journal of the American Medical Association* User's Guide to the Medical Literature series[11]), none have been developed for OERs.

Several recent studies have explored how to evaluate blogs and podcasts. Using a modified systematic review, Paterson et al. found 151 quality indicators for secondary resources in the existing educational literature that may be relevant for these resources.[12] Subsequently, medical educators in various specialties as well as expert bloggers and podcasters in emergency medicine and critical care endorsed many of these quality indicators in two modified Delphi studies.[13-14] Another rating tool, dubbed the *Academic Life in Emergency Medicine* Approved Instructional Resources (ALiEM AIR) Score, was developed for use by groups of medical educators.[15] This score was based on a best approximation of what educators thought were key features of a robust blog post or podcast summary. None of these studies, however, provided a practical, simplified scoring tool to help health professionals and trainees assess the quality of OERs.

In this study, we attempted to translate the information from the previous review of the literature[12] and modified Delphi studies[13-14] to create a functional quality evaluation instrument to guide trainees in critical appraisal of blog or podcast-related written materials.

METHODS

This study was conducted in three phases. Phase 1 reduced a previously derived and evaluated list of quality indicators to a manageable number for further assessment using data reduction techniques. Phase 2 further evaluated the remaining quality indicators in a group of trainees. We used these data to derive quality evaluation instruments and assess their reliability. Phase 3 assessed the concordance of the derived instruments with two currently accepted methods of quality evaluation (ALiEM AIR certification and educator gestalt).

An institutional review board granted an exemption for all three phases of the study. Phase 1 of the study involved the further analysis of data obtained in three previous studies[13, 16-17] that were granted exemptions by the Hamilton Research Ethics Board (http://fhs.mcmaster.ca/healthresearch/hireb.html). Phase 2 and 3 also received an exemption. Phases 2 and 3 involved a multi-centre, web-based, cohort rating study that was conducted during April-August 2015.

Phase 1: Quality Indicator Selection.

This study built upon the work of three previously published studies. Paterson et al. defined 151 potential quality indicators that could be applied to OERs such as blogs and podcasts.[12] This extensive list, however, is too unwieldy for learners to use practically in guiding their decision-making for appraising OERs. Subsequently, two consensus-building Delphi studies were conducted to identify what expert groups (medical educators, expert podcasters, and expert bloggers) felt were the most important quality indicators.[13-14, 18] For the purposes of Phase 1 of this study, iterative steps were made to shorten the list of quality indicators.

The overall process is depicted in Figure 1. First, we examined the priorities of expert groups (medical educators, expert OER producers) from two previous modified Delphi studies.[13-14] These expert groups were selected by peer nomination via snowball sampling technique[14] or by self-determination through attendance at an international consensus conference.[13] In both of these studies all 151 items were ranked on seven-point Likert scales (1=strongly disagree, 7=strongly agree with item). As such, we were able to use these data to calculate item total correlations (ITC) for the 151 possible quality indicators. ITCs are an indication of the relationship between individual items and the measurement of the scale. We eliminated items with an ITC of less than 0.3, because low ITCs can be used to eliminate items that poorly fit with the scale's measurement construct.[19]

Items with a low mean score across all the experts in the two Delphi groups (i.e. rated <5.5 on the 7-point scale) were also eliminated as possible items for our score derivation. To ensure that we valued the ratings of all groups, we also conducted a principle component analysis to look at the groupings of priorities across the groups of educators, podcasters, and bloggers.

Finally, we conducted a two-round consensus building exercise within our study team's clinician educators (TC,

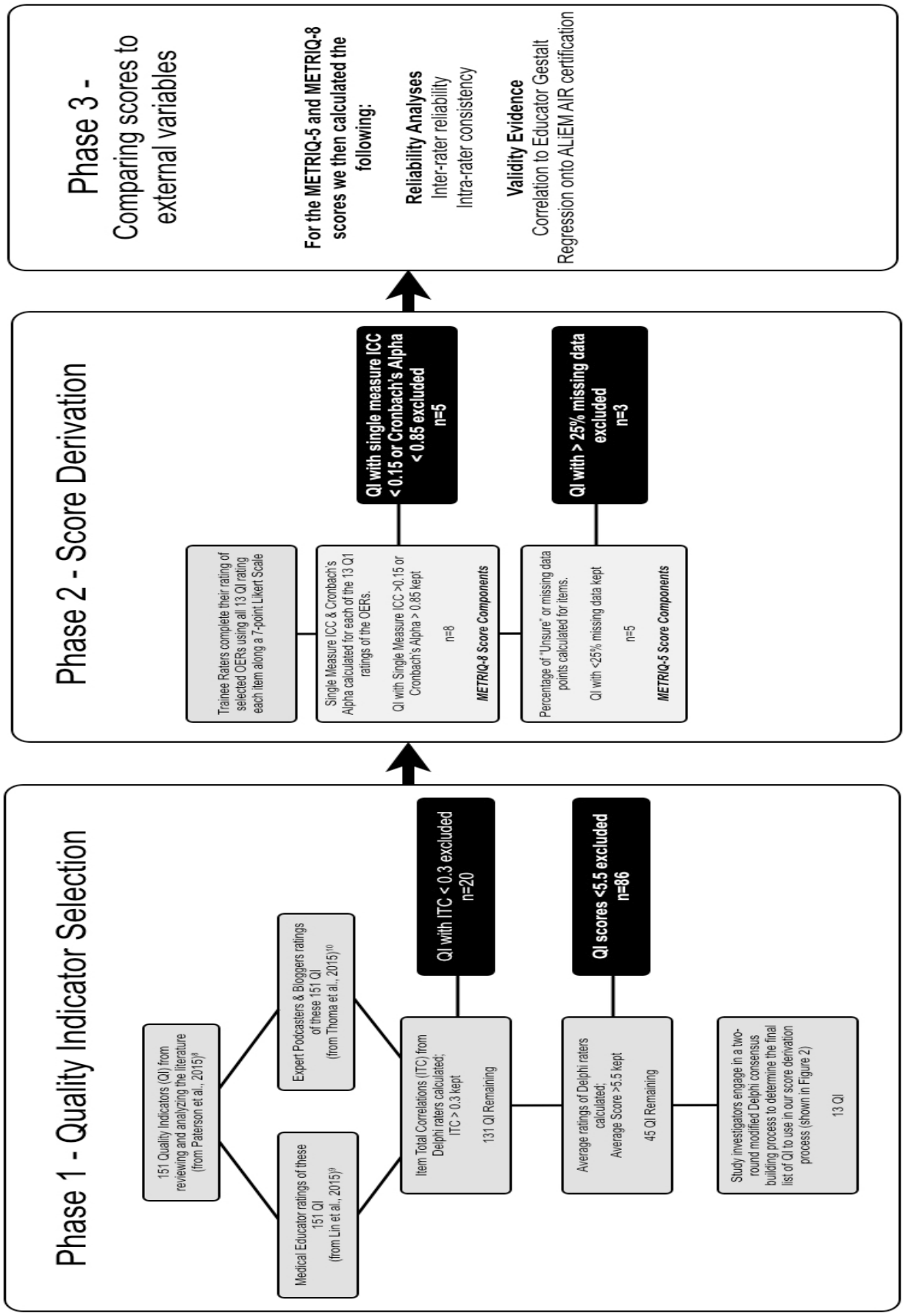

Figure 1. Flow diagram of study design. Phase 1 depicts the quality-indicator (QI) selection process, Phase 2 depicts the score derivation process based on the reduced list of QIs, and Phase 3 describes the reliability and validity testing data for the two derived instruments for scoring the quality of medical blogs and podcasts.
QI, quality indicator; *ITC*, item total correlation; *ICC*, intraclass correlation coefficient; *ALiEM*, Academic Life in Emergency Medicine; *AIR*, approved instructional resources.

Table 1. Parent websites and distribution of the 39 selected blog or podcast online educational resources (OER), from which the gestalt score was derived (Phase 2).

Website name	Number of rated posts
Academic Life in Emergency Medicine	12
BoringEM	1
Clinical Monster	1
Dr. Smith's ECG blog	2
Don't Forget The Bubbles	2
Emergency Medicine Ireland	1
EM Lyceum	3
EM Basic	1
EMCrit	1
EM Literature of Note	1
ERCast	3
Life in the Fast Lane	2
Pediatric EM Morsels	4
R.E.B.E.L EM	1
The NNT	1
The Poison Review	2
The Skeptics Guide to Emergency Medicine	1

NB: For a complete listing of all the rated blog posts, please refer to Appendix.

BT, ML, CC, MA) to determine items we felt would be most easily rated by junior learners without training. Our team focused on eliminating items that demonstrated any of the following: required extensive knowledge or expertise, were difficult to judge without training, or were difficult to understand or define.

Phase 2: Critical Appraisal Score Derivation

Rater Population and Materials. Participating collaborators were trainees (medical students, n=36; residents, n=9) from Canada and the United States, who were recruited from centers affiliated with our investigatory team and by a snowball referral process. The participants are all listed as collaborators in this study in the acknowledgments section and participated voluntarily.

The rated materials were drawn from a list of openly accessible online blog posts, previously rated for educational merit by the ALiEM AIR program (http://www.aliem.com/new-air-series-aliem-approved-instructional-resources/).[15] From a list of the initial 80 ALiEM AIR-rated OERs, we randomly selected 39 (20 were ALiEM AIR certified as good quality, and 19 that were not) for inclusion in Phase 2. Table 1 lists the parent websites for these 39 blog post or podcast-related OERs, and Appendix lists each OER's website addresses and expert gestalt ratings.

Data Collection and OER Scoring. Participating trainee raters were given three months to rate 39 OERs using a web-based Google Forms survey. Each OER was rated on 13 potential scoring system items from our reduced list (Figure 2). Each item was rated upon a seven-point Likert scale, which was anchored at 1 by the statement "Attribute not displayed," and at 7 by the statement "Attribute displayed well."

One OER was rated twice by each rater to allow for a calculation of intra-rater consistency. We used a modified Dillman technique to provide raters with three reminders over the study duration.[20]

Derivation of Our Scoring System Models. To derive our proposed scoring systems, we calculated the single measure intraclass correlation coefficient (ICC) and Cronbach's alpha for each of the 13 potential scoring system items for all the trainee-rated OERs.[19] We also calculated a repeated-measures ANOVA to determine intra-rater consistency for the 13 potential subscores (quality indicators) for the same rater rating the same OER at two different times.

As there were many 'missing data' due to rater uncertainty, we used the imputation model of substituting the grand mean for each quality indicator item to compensate for these. This imputation technique is deemed a highly conservative approach for calculating an ICC and Cronbach's alpha. A subset of the investigatory team (TC, KK) then set a Cronbach's alpha threshold (or average measures ICC) of ≥0.85 and a Single Measure ICC of ≥0.15 in order to derive our first scoring system model, as we felt that items that scored <0.15 in the ICC would be considered quite poor. Of note, single measure ICC measures of 0.1-0.2 are considered poor, 0.3-0.4 are considered fair, 0.5-0.6 considered moderate, 0.7-0.8 indicates strong agreement, and >0.8 indicates almost perfect.[19] The items that met these thresholds were used to generate the first model.

Table 2. Educator gestalt rating scale of blogs and podcasts for trainee learning.

Would you recommend this to a learner?							
0 Unsure	1 No, this is an inappropriate resource for this audience	2	3	4 This may be useful to this audience	5	6	7 Yes, this is a great resource for this audience

Q1. Universal technology - Does the resource employ technologies that are universally available to allow learners with standard equipment and software access?
Q2. Maintenance - Is the resource maintained such that its text and multimedia elements remain functional?
Q3. Concise content - Does the resource contain an appropriate amount of information for its length?
Q4. Scholarly use of language - Does the resource use efficient, accurate language that is appropriate for its target audience?
Q5. Is the editorial process independent from sponsors, conflict of interest, and other sources of bias?
Q6. Are the processes (e.g. editorial, peer review, evaluation, etc) that were used to create the resource outlined?
Q7. References - Does the resource cite its references?
Q8. Editorial process - Is there an editorial process?
Q9. Consistency with citations - Are the resource's statements consistent with its references?
Q10. Background - Does the resource provide enough background information to situate the learner in the context of prior knowledge?
Q11. Moderation - Are interactions between learners moderated effectively to ensure professional conduct?
Q12. Publisher - Is it clear who published the resource?
Q13. Reading/Listening - Is the resource composed in a way that makes it easy to understand? (not overly convoluted)

Figure 2. Final list of 13 quality indicators rated by trainee raters on a 7-point Likert scale.

Our second model incorporated the previous model, but eliminated items that generated a substantial amount of missing data (i.e. rated as "unsure"). For practicality, we felt it was important for individual raters to be able to use the quality indicator subscore items. Therefore, any items yielding a substantive amount of missing data (i.e. >25% of items were unable to be scored by the trainee raters) were eliminated as well.

Phase 3: Comparing the scoring models with educator gestalt and ALiEM AIR ratings

Rater Population, Materials, and Data Collection. Participating collaborators for educator gestalt ratings were practicing academic emergency physician volunteers with a primary interest in medical education (n=20) from Canada and the United States. The participants were recruited by members of the investigatory team (TC, BT, ML, CC, MA) and are all listed as collaborators in this study in the acknowledgments section. ALiEM AIR certification status information was taken from the first six modules listed on the ALiEM.com webpage (https://www.aliem.com/aliem-approved-instructional-resources-air-series/).[14]

Outcome Variables:

Other Critical Appraisal Methods. Informed by the components of external validity described by Messick,[22] we

Table 3. Demographics of raters who evaluated online educational resources.

	Instrument development trainee raters (n=40)		Expert gestalt educator raters (n=20)	
% by country of origin	2.5% United States of America 97.5% Canada		75% United States of America 25% Canada	
Year of training or years in practice at the time of their enrollment	0 years in practice (All are trainees)		10.3 years in practice (SD 10.2)	
Academic affiliation	Year 1 medical student	40%	Full professor	10%
	Year 2 medical student	30%	Associate professor	15%
	Year 3 medical student	18%	Assistant professor	65%
	Year 4 medical student	3%	Clinical appointment	10%
	Year 1 resident	5%	None	5%
	Year 2 resident	3%		
	Year 3 resident	3%		
% current or past official medical education position within institution	N/A		90% total Breakdown	
			Dean / chair	15%
			Residency PD	40%
			Residency APD	45%
			Other GME role	30%
			Clerkship director / UGME role	30%
			Research/quality Improvement role	20%

PD, program director; *APD*, associate or assistant program director; *GME,* graduate medical education; *UGME,* undergraduate medical education

Table 4. Correlations between the scores by subjects in the first and second rating incidence.

Question item number	Pearson's *r* between the first rating and second rating of each possible quality indicator subscore item	p-value
Q1	0.92	<0.001
Q2	0.84	<0.001
Q3	0.37	0.05
Q4	0.63	<0.001
Q5	0.33	0.08
Q6	0.45	0.02
Q7	0.93	<0.001
Q8	0.57	0.001
Q9	0.74	<0.001
Q10	0.71	<0.001
Q11	0.79	<0.001
Q12	0.81	<0.001
Q13	0.85	<0.001

compared the scoring models to other existing measures of quality for OERs.

The 39 trainee-scored OERs were rated by educators using the same data collection method outlined in Phase 2. However, rather than rating each OER using the 13 quality indicators, the faculty were asked to use their gestalt, expert judgment to decide whether the OER would be acceptable for trainee learning. See Table 3 for the qualifications of the faculty raters. Educator's gestalt was rated using a seven-point Likert scale (Table 2).

In addition to the educator gestalt score, the ALiEM AIR certification process served as another comparative scoring system. This was a separate rating process external to our study and raters with a separate panel of nine expert faculty panellists selecting OERs for a resident audience. The certification of these posts is openly accessible via the

Table 5. Inter-rater agreement on the quality indicator subscore components, calculated using a 2-way random effects model for consistency to calculate the ICCs (interclass correlation coefficient).

Question item number	Single measure ICC*** (95% CI)	Average measure ICC*** (95% CI)	Number of missing data points	% Missing
Q1*	0.04 (0.02-0.08)	0.64 (0.47-0.79)	202	13%
Q2*	0.03 (0.01-0.07)	0.56 (0.35-0.74)	193	12%
Q3	0.17 (0.12-0.26)	0.89 (0.84-0.94)	206	13%
Q4*	0.12 (0.07-0.19)	0.84 (0.76-0.90)	208	13%
Q5*	0.10 (0.06-0.16)	0.81 (0.71-0.89)	713	45%
Q6**	0.28 (0.20-0.39)	0.94 (0.91-0.96)	476	30%
Q7	0.38 (0.28-0.50)	0.96 (0.94-0.98)	216	14%
Q8**	0.22 (0.15-0.32)	0.92 (0.89-0.95)	773	48%
Q9**	0.16 (0.11-0.25)	0.88 (0.82-0.93)	465	29%
Q10	0.22 (0.14-0.32)	0.92 (0.87-0.95)	287	18%
Q11	0.17 (0.11-0.26)	0.89 (0.83-0.93)	290	18%
Q12	0.29 (0.21-0.41)	0.95 (0.92-0.97)	319	20%
Q13*	0.14 (0.09-0.22)	0.87 (0.80-0.92)	285	18%

* Eliminated in Score Models 1 and 2 due to alpha <0.85 or single measure ICC <0.15

** Eliminated in Score Model 2 since trainees were unsure too often (>25% missing data)

*** p-value was <0.001 for all ICC calculated

Score Model 1: METRIQ-8 Score (Maximum 56 points)	Score Model 2: METRIQ-5 Score (Maximum 35 points)
Q3 Concise content - Does the resource contain an appropriate amount of information for its length?	Q3 Concise content - Does the resource contain an appropriate amount of information for its length?
Q6 Content Construction - Are the processes (e.g. editorial, peer review, evaluation, etc) that were used to create the resource outlined?	Q7 References - Does the resource cite its references?
Q7 References - Does the resource cite its references?	Q10 Background - Does the resource provide enough background information to situate the learner in the context of prior knowledge?
Q8 Editorial Process - Is there an editorial process?	Q11 Moderation - Are interactions between learners moderated effectively to ensure professional conduct?
Q9 Consistency with citations - Are the resource's statements consistent with its references?	Q12 Publisher - Is it clear who published the resource?
Q10 Background - Does the resource provide enough background information to situate the learner in the context of prior knowledge?	
Q11 Moderation - Are interactions between learners moderated effectively to ensure professional conduct?	
Q12 Publisher - Is it clear who published the resource?	

Figure 3. Two proposed online educational resources evaluation instruments.

Internet.[21] Of note, those who had acted as an ALiEM AIR rater were excluded from rating for this present study.

Validity Evidence

Akin to many clinical decision rule (CDR) study designs, we opted to perform regression analyses using our two newly derived score models to determine whether they would regress to two comparative scoring instruments: the educator gestalt score and the ALiEM AIR certification using a binary logistic regression model. For the purposes of the correlation analyses, we chose to use the pragmatic score models (with substitution of a zero score when there were missing data) since individual users would not have access to grand means for the subscore components.

RESULTS

Phase 1: Quality Indicator Selection

The overall results and process are depicted in Figure 1. ITCs for the 151 possible quality indicators were calculated using data from the previous Delphi studies.[13-14] Twenty items had an ITC<0.3, and 81 of the remaining items were rated <5.5 on the seven-point Likert scale across the two Delphi groups, and thus they were eliminated. The two-round, consensus-building exercise within our study team identified 13 of the final 45 items as being most easily rated by trainees. This list is outlined in Figure 2.

Phase 2: Score Derivation

Table 3 depicts the demographics for the 60 total volunteers, who were recruited for the OER rating exercises.

Of this group, 28 of the 40 trainee raters (27 medical students, one resident) completely reviewed all OERs in our study. The remaining 12 trainee raters yielded incomplete datasets requiring the use of an imputation model to calculate the ICC in our score derivation procedures as described in the methods section. All 20 educators generating the gestalt ratings reviewed the complete set of OERs.

Intra-Rater Consistency for the 13 Quality Indicators

Since one item was rated at two different points in our

Table 6. A comparison of the reliability calculations of the two proposed online educational resources evaluation instruments using different missing data procedures.

	METRIQ-8 score		METRIQ-5 score	
	Pragmatic analysis	Imputation analysis	Pragmatic analysis	Imputation analysis
Single measure ICC (95% CI)	0.30 (0.22-0.42)	0.38 (0.29-0.51)	0.22 (0.15-0.32)	0.35 (0.26-0.47)
Average measure ICC (95% CI)	0.94 (0.92-0.97)	0.96 (0.94-0.98)	0.92 (0.88-0.95)	0.96 (0.93-0.97)

ICC, intraclass correlation coefficient

*NB: The pragmatic analysis awards a zero value to any missing data points. The imputation analysis substitutes the grand mean for the missing data points (any items which were not rated by the trainee raters).

Table 7. Relationships between average METRIQ-8 and METRIQ-5 Scores with other comparative instruments (average educator gestalt score, ALiEM AIR certification).

	METRIQ-8 score pragmatic score	METRIQ-5 score pragmatic score
Pearson correlation (r) to educator gestalt score for recommending resource to a trainee	r=0.35 p=0.03	r=0.41 p<0.01
Logistic regression for ALiEM AIR certification status	Odds ratio 1.28 (1.09-1.50) Wald test (1,38)=8.8 p=0.003	OR = 1.5 (1.14-2.20) Wald test (1,38)=8.4 p=0.004

rating exercise by our trainee raters, we were able to calculate a measure of internal consistency for the various items. For this analysis, we eliminated raters with incomplete data sets, using only the remaining raters to calculate a repeated-measures ANOVA to determine if there was a significant change in the quality indicator subscores when the rater encountered the OER on the second occasion. We did not detect a significant main effect of the repeated measurement occasion in our analysis ($F=0.54$, df (1), $p=0.47$). Across the 13 conditions, the first and second ratings of this item mostly correlated. We calculated the Pearson correlations for these scores, which ranged from 0.33 to 0.93 for the various items (Table 4).

Inter-Rater Reliability for the 13 Quality Indicators

After applying our selected imputation model (substitution of grand mean) to compensate for missing data, we calculated the intraclass correlation coefficients for each of the 13 quality indicator subscores. We used two-way random effects model for consistency measures to determine the single and average measure ICCs (Table 5). A single measure ICC allows us to understand the consistency of a randomly drawn single rater's scores. The average measure ICC gives the reliability of the score generated by averaging or totalling the scores of *all the raters* who evaluated the OER. It can help estimate how reliability is improved by increasing the number of raters or ratings and give an indication of the actual reliability of the score generated by using several raters.[19] This eliminated five of our possible quality indicator subscores items to generate the eight-item Score Model 1.

Missing Data Across the 13 Quality Indicator Subscores

Certain items yielded a high number of missing data points because participants were unsure whether to rank these items. For the purposes of deriving the score, we felt it would be prudent to generate a score model that only included items with a low number of missing data points. We therefore used a cut off of >25% missing data points within a subscore dataset to eliminate another three items from the list in Score Model 1 (eight items) to generate Score Model 2 (five items).

Properties of the Scores

Score Model 1 and 2 propose an eight-component and five-component score, respectively, which we will hereafter refer to as the METRIQ 8 Score and METRIQ 5 Score, respectively. Figure 3 lists the subscores for both OER evaluation instruments, proposed by this derivation study.

Reliability of the Aggregate Scores for METRIQ-8 and METRIQ-5

For the reliability calculation of the aggregate scores, we used both a pragmatic analysis which included 0-scores for any facet where a trainee rater was unsure and also an imputation analysis which included the grand mean of the subscore item. Both models were found to be moderately reliable regardless of the analytic approach with $p<0.001$, with the METRIQ-8 performing slightly more reliably than METRIQ-5. (Table 6).

Phase 3: Comparing the scoring models with educator gestalt and ALiEM AIR ratings

We evaluated our scoring model instruments against both educator gestalt and ALiEM AIR certification status. We first determined the correlation between our METRIQ-8 and METRIQ-5 models and average educator gestalt score for 20 educators. We also used a logistic regression model to determine if our models would regress upon the ALiEM AIR certification status (certified or not).

Correlation Between Mean Educator Gestalt Score and the Average METRIQ-8 and METRIQ-5 Scores

To strengthen the validity evidence for our nascent scoring systems, we calculated the Pearson correlation statistic for the average educator gestalt scores and the pragmatic versions of both METRIQ-8 and METRIQ-5. We detected moderate correlations ($p<0.05$ for both) between our proposed scores and the average educator gestalt scores as shown in Table 7.

Logistic Regression onto ALiEM AIR Certification Status

To determine if our score had a relationship with ALiEM AIR certification, we conducted a binary logistic regression on the ALiEM AIR certification status. As demonstrated by the Wald test, this yielded a significant odds ratio for both scores. The odds ratios for METRIQ-5 and METRIQ-8 scores were 1.28, ($p=0.03$) and 1.5 ($p=0.004$) respectively.

DISCUSSION

Teaching clinical providers the skill of critically appraisal OERs will be increasingly important as blogs and podcasts proliferate.[4] With traditional secondary resources such as textbooks and lectures, the credibility of the source of these teachings (i.e. the editorial board of a textbook or the professorial status of a teacher) are often cited as the rationale behind why trainees and educators accept these resources as unequivocally valid without formal critical appraisal. While neither trainees nor educators have traditionally given much thought to the critical appraisal of these traditional secondary resources, the ubiquity and accessibility of OERs makes it imperative that we begin to teach trainees to be both judicious and educated in their use of these resources. Similar to what the DISCERN score did for online patient-oriented materials,[23-24] our proposed METRIQ-8 and METRIQ-5 scores may allow us to ensure that trainees and educators are better able to appraise the quality of the resources they use to learn and teach, respectively.

Our investigatory team derived two scoring systems by drawing on the tradition of creating clinical decision rules (CDRs) to guide novice decision-making in patient care. We have attempted to follow a rigorous derivation process in this study, akin to those used to derive CDRs.[25-26] In fact, the culmination of this study is equivalent to a Level 4 derivation study.[26] Both of the proposed evaluation scoring instruments will require external validation. The METRIQ-8 score performs slightly better in terms of reliability. Its higher reliability may be a result of purely having more items, and thus yielding greater precision. In contrast, the METRIQ-5 score may be more easily used by trainees given its brevity (only five questions) and decreased complexity. The METRIQ-5 score may correlate better with other external measures of quality for these reasons.

Moving forward, further testing of the METRIQ scores in various populations will be required as reliability and validity are context specific, and depend on how the scores are used. METRIQ-8 and METRIQ-5 will need to be evaluated by separate and internationally diverse rater populations to provide further validity evidence, support their use, and extend their generalizability. Additionally, head-to-head comparisons with other scoring systems (such as the ALiEM AIR score, which is meant to be used by faculty members when selecting educational resources) will be necessary.[15] We were only able to look at the relationship of our new scores with ALiEM AIR certification status (i.e. awarded or not). The use of this dichotomous data (certified or not) rather than the detailed score results (a continuous score ranging from 0 to 35) may have limited our calculations. Finally, a prospective study design looking at whether these instruments correlate with usage (i.e. webpage views or social media sharing) may be useful.

In a previous study by our research group, we found that trainees were able to select resources with single-measure ICCs of 0.22 for each other.[10] The use of the pragmatic METRIQ-8 score improves upon this while the METRIQ-5 score approximates this consistency but further defines what may guide that gestalt. The much higher average measures ICCs suggest that a group-based rating system may be best for selection of resources for trainees. Much akin to other crowd-based rating systems (e.g. BEEM rating score[27-28] and Yelp), group-based decision-making ultimately may be the best guide for rating individual resources.

LIMITATIONS

There are several major limitations to this study. First, the use of the medical educator gestalt score as a reference standard may be questionable, since this measure has been shown to be insufficiently reliable and lacking sufficient validity evidence to provide consistent guidance to trainees.[10] However, it is the most commonly used method for determining the quality of OERs. Second, we have used uncalibrated raters. Previous research has shown that rater cognition improves significantly if we use calibration processes such as rater-training.[29] Third, we used a convenience sampling of raters in both the trainee and medical educator groups, which may have been biased by their contact with our investigatory group, although we attempted to sample broadly from multiple centres. We are actually quite hopeful that with rater training and calibration the use of the METRIQ scores could be improved. Fourth, our methods may be critiqued for being overly complicated. We have attempted to use robust and reproducible methods for reducing the 151 possible quality indicators that were previously found in the literature.[12] In an effort to aggressively reduce this list, we used fairly novel methods to create two sensibly compact evaluation instruments that may be reliably applied by trainees. As such, it is prudent to compare our new scores directly with other known scores such as the ALiEM AIR before extensive use. Moreover, this study also attempts to gather some validity evidence to support the two proposed scores, but is limited because we used the non-blinded ALiEM AIR certification status of OERs to compare with our two proposed scoring instruments. Finally, many of the authors for this paper are website editors, authors, or affiliated in some way with the various blogs listed used for this study. To minimize the effects of our bias, we sought collaborators with fewer stakes and affiliations (i.e. the peer-nominated experts) to review the materials. We also included members of the team (CC, KK, KK) who are not significantly invested in these OER outlets to provide some level of objectivity and reflexivity to our investigator team.

CONCLUSION

We have derived two possible evaluation instruments (METRIQ-8 and METRIQ-5), which may help trainees identify higher quality OERs, establish a precedent for reviewing and critically appraising secondary resources, and guide OER producers (bloggers and podcasters) to improve the quality of their educational content. These instruments correlated favourably with experienced faculty educator gestalt ratings of online educational resources.

ACKNOWLEDGEMENTS

We would like to thank all our study collaborators who acted as trainee raters for this study:

Amanpaul Bhamber (McMaster University), Laura Bosco (Queen's University), William Caron (McMaster University), Graham Chamberlain (McMaster University), Dr. Philip Chan (Washington University in St. Louis), Dr. Rory Connolly (University of Ottawa), Emily Dewhurst (McMaster University), Isabelle Dobronyi (McMaster University), Justina Ellery (McMaster University), Larissa Hattin (McMaster University), Dr. Ariel Hendin (University of Ottawa), Abid Ismail (McMaster University), Michelle Kuang (McMaster University), Ryann Kwan (McMaster University), Eric Lachance (University of Manitoba), Rebecca Lang (University of Manitoba), Ian Laxdal (University of Manitoba), Meirui Li (McMaster University), Sherry Liu (McMaster University), Gordon Locke (McMaster University), Jennifer McCall (Queen's University), Chris Meyer (McMaster University), Adam Mutsaers (McMaster University), Sheena Nandalal (McMaster University), Laila Nasser (McMaster University), Anna Naylor (McMaster University), Sam Neily (University of Manitoba), Melanie Ngo (University of Manitoba), Taylor Oliver (University of Saskatchewan), Quinten Paterson (University of Saskatchewan), Laura Pellow (McMaster University), Beatrice Preti (McMaster University), Nina Ramic (Queen's University), Priya Sharma (University of Manitoba), Tara Stratton (McMaster University), Dr. Rajiv Thavanathan (University of Ottawa), Jenifer Truong (McMaster University), Alex Van Duyvendyk (University of Saskatchewan), Molly Whalen-Browne (McMaster University), Paul Young (McMaster University). We would like to thank the following faculty collaborators who lent their expertise to us as faculty raters: Dr. James Ahn (University of Chicago); Dr. Matt Astin (Mercer University); Dr. Robert Cooney (Geisinger Health Systems); Dr. Sandy Dong (University of Alberta); Dr. Justin Hensley (Texas A&M/ Christus Spohn); Dr. David Jones (Oregon Health Sciences University) Dr. Ian Julie (University of California at Davis); Dr. Gloria Kuhn (Wayne State University); Dr. Sean Moore (University of Ottawa); Dr. Erik Nordquist (Cook County Health and Hospital System); Dimitri Papanagnou (Sidney Kimmel Medical College at Thomas Jefferson University); Dr. Alim Pardhan (McMaster University); Dr. Hans Rosenberg (University of Ottawa); Dr. Christopher Ross (Cook County Health and Hospital System); Dr. Mike Schindlbeck (Cook County Health and Hospital System); Dr. Mark Silverberg (Kings County); Dr. Nelson Wong (Mount Sinai, NY); Dr. Andrew Worster (McMaster University); Dr. Brian Wright (Stony Brook).

Address for Correspondence: Teresa Chan, MD, MHPE, McMaster University, Department of Medicine, Division of Emergency Medicine, Mc Master Clinic, 237 Barton Street East, Room 254, Hamilton Ontario, Canada, L8L 2X2. Email: teresa.chan@medportal.ca.

Conflicts of Interest: By the *West*JEM article submission agreement, all authors are required to disclose all affiliations, funding sources and financial or management relationships that could be perceived as potential sources of bias. Drs. Brent Thoma and Teresa Chan are on the volunteer editorial team of CanadiEM/ BoringEM blogs. Drs. Brent Thoma, Michelle Lin, and Teresa Chan are all volunteer editorial members of the ALiEM blog.

REFERENCES

1. Wilinsky J. The unacknowledged convergence of open source, open access, and open science. *First Monday*. 2005;10(8).
2. Suber P. Timeline of the open access movement. 2009. Available at: http://legacy.earlham.edu/~peters/fos/timeline.htm.
3. Nickson CP, Cadogan MD. Free Open Access Medical education (FOAM) for the emergency physician. *Emerg Med Australas*. 2014;26(1):76-83.
4. Cadogan M, Thoma B, Chan TM, et al. Free Open Access Meducation (FOAM): the rise of emergency medicine and critical care blogs and podcasts (2002-2013). *Emerg Med J*. 2014;e1:e76-e77.
5. Mallin M, Schlein S, Doctor S, et al. A survey of the current utilization of asynchronous education among emergency medicine residents in the United States. *Acad Med J Assoc Am Med Coll*. 2014;89(4):598-601.
6. Purdy E, Thoma B, Bednarczyk J, et al. The use of free online educational resources by Canadian emergency medicine residents and program directors. *Can J Emerg Med*. 2015;17(2):101-6.
7. Pearson D, Bond M, Kegg J, et al. Evaluation of Social Media Use by Emergency Medicine Residents and Faculty. *West J Emerg Med*. 2015;16(5):715-20.
8. Green ML. Graduate Medical Education Training in Clinical Epidemiology, Critical Appraisal, and Evidence-based Medicine: A Critical Review of Curricula. *Acad Med*.1999;74:686-94.
9. CACMS STANDARDS AND ELEMENTS Standards for Accreditation of Medical Education Programs Leading to the MD Degree. 2015. Available at: https://www.afmc.ca/pdf/CACMS_Standards_and_Elements_June_2014_Effective_July12015.pdf.
10. Krishnan K, Trueger NS, Thoma B, et al. Gestalt Assessment of

Online Educational Resources is Unreliable and Inconsistent. In: Canadian Conference on Medical Education, Accountability from Self to Society. Montreal, QC, Canada; 2016.

11. Guyatt G, Rennie D, Meade M, et al. Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice, Third Edition. 3rd ed. McGraw-Hill; 2015.
12. Paterson QS, Thoma B, Milne WK, et al. A Systematic Review and Qualitative Analysis to Determine Quality Indicators for Health Professions Education Blogs and Podcasts. *J Grad Med Educ*. 2015;7(4);549-54.
13. Lin M, Thoma B, Trueger NS, et al. Quality indicators for blogs and podcasts used in medical education: modified Delphi consensus recommendations by an international cohort of health professions educators. *Postgrad Med J*. 2015;91(1080):546-50.
14. Thoma B, Chan TM, Paterson QS, et al. Emergency Medicine and Critical Care Blogs and Podcasts: Establishing an International Consensus on Quality. *Ann Emerg Med*. 2015.
15. Lin M, Joshi N, Grock A, et al. Approved Instructional Resources (AIR) Series: A national initiative to identify quality emergency medicine blog and podcast content for resident education. *J Grad Med Educ*. 2016;8(2):219-25.
16. Thoma B, Sanders JL, Lin M, et al. The social media index: measuring the impact of emergency medicine and critical care websites. *West J Emerg Med*. 2015;16(2):242-9.
17. Paterson QS, Thoma B, Lin M, et al. Quality Indicators for Medical Education Blog Posts and Podcasts: A Qualitative Analysis and Focus Group. In: Association of American Medical Colleges Medical Education Meeting. Chicago; 2014.
18. Paterson QS, Colmers IN, Lin M, Thoma B, Chan T. The quality checklists for health professions blogs and podcasts. 2015:1-7.
19. Norman GR and Streiner DL. Health Measurement Scales: A Practice Guide to Their Development and Use. Third. Oxford, UK: Oxford University Press; 2008.
20. Dillman D. Mail and Internet Surveys: The Tailored Design Method. 2nd ed. Hoboken, NJ: John Wiley & Sons; 1999.
21. ALiEM Approved Instructional Resources (AIR Series). Available at: ALiEM.com. http://www.aliem.com/aliem-approved-instructional-resources-air-series/. Accessed Jan 11, 2016.
22. Messick S. Validity. In: Linn RL, ed. Educational Measurement. Third. Macmillan; 1989.
23. Rees CE, Ford JE, Sheard CE. Evaluating the reliability of DISCERN: A tool for assessing the quality of written patient information on treatment choices. *Patient Educ Couns*. 2002;47:273-5.
24. Charnock D, Shepperd S, Needham G, et al. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health*. 1999;53(2):105-11.
25. Stiell IG and Wells G a. Methodologic Standards for the Development of Clinical Decision Rules in Emergency Medicine. *Ann Emerg Med*. 1999;33(4):437-47.
26. Mcginn TG, Guyatt GH, Wyer PC, et al. Users ' Guides to the Medical Literature XXII : How to Use Articles About Clinical Decision Rules. 2015;284(1):79-84.
27. Worster A, Kulasegaram K, Carpenter CR, et al. Consensus conference follow-up: inter-rater reliability assessment of the Best Evidence in Emergency Medicine (BEEM) rater scale, a medical literature rating tool for emergency physicians. *Acad Emerg Med*. 2011;18(11):1193-1200.
28. Carpenter CR, Sarli CC, Fowler S a, et al. Best Evidence in Emergency Medicine (BEEM) rater scores correlate with publications' future citations. *Acad Emerg Med*. 2013;20(10):1004-12.
29. Kogan JR, Conforti L, Bernabeo E, et al. Opening the black box of clinical skills assessment via observation: A conceptual model. *Med Educ*. 2011;45(10):1048-60.